A Note on Shumailov et al. (2024): 'AI Models Collapse When Trained on Recursively Generated Data'

Year: 2024 | Citations: 104 | Authors: Ali Borji

Abstract

The study conducted by Shumailov et al. (2024) demonstrates that repeatedly training a generative model on synthetic data leads to model collapse. This finding has generated considerable interest and debate, particularly given that current models have nearly exhausted the available data. In this work, we investigate the effects of fitting a distribution (through Kernel Density Estimation, or KDE) or a model to the data, followed by repeated sampling from it. Our objective is to develop a theoretical understanding of the phenomenon observed by Shumailov et al. (2024). Our results indicate that the outcomes reported are a statistical phenomenon and may be unavoidable.